BIOLOGICAL CRITERIA
Technical Guidance for Survey Design and Statistical Evaluation of Biosurvey Data

CHAPTER 5 Discussion and Examples

In the previous four chapters, standard statistical methods were presented, discussed, and illustrated with simple examples. Those methods and examples represent conventional analyses or situations in which sample sizes are relatively large so that hypothesis testing is essentially straightforward. The analyses were motivated by available, commonly applied methods, and the examples were structured to fit the methods. The purpose was to provide background statistical guidance, with examples.

In this chapter a different approach is taken. Here, typical problems involving biosurvey data are the starting points, and statistical methods for analysis and hypothesis testing are proposed and applied specifically to the problem. In some cases, hypothesis testing is possible; in others, the small sample size may limit statistical inference. In the latter situation, the investigator may consider design changes so that different statistical analyses can be undertaken with biosurvey data in the future.

We begin with a general discussion of the importance of small sample size and briefly examine judgmental and statistical options for small sample size, followed by examples of hypothesis testing with small samples. The chapter concludes with "rules of thumb."

Working with Small Sample Sizes

The conventional methods for statistical hypothesis testing and interval estimation presented in chapters 1 through 4 work best under conditions that do not always exist with biosurvey data. The common approaches based on an underlying normal probability model are clearly not essential; distribution-free methods are versatile and effective. Still, virtually all confirm-atory analyses (i.e., those concerned with hypothesis testing and interval estimation) require estimation of a "location" statistic that is the quantity of interest (e.g., a mean, median, or quartile), and they also require estimation of a variability statistic (e.g., a standard error) that indicates the spread of values for the location statistic.

An example of a desirable scenario for confirmatory statistical analysis was described in Chapter 2. Data must be available from the sites of direct interest in the assessment, and sample sizes must be large enough for hypothesis testing. If the site-specific data are inadequate (less than two, which would prevent direct calculation of a sample variance), or too small, (e.g., less than five, which would make the calculated sample variance quite uncertain), then alternatives to statistical testing or intervals are possible, but these alternatives are apt to include additional conditions or assumptions beyond those required in conventional analyses.

For example, a single sampling might yield a point estimate for IBI downstream of a wastewater discharge, but provide no measure of variability. If historic data exist on IBI at other impacted sites, then it is reasonable to assume that the variability in the historic data can be used as the variability measure for testing at the site of interest. If, on the other hand, the historic data analysis includes an IBI regression based on predictors, such as watershed area and physical habitat quality, then the standard error for this regression is the appropriate variability measure. The key feature of these hypothetical examples is that other, relevant information exists that the investigator believes can be used to estimate statistics for the site of interest.

In the absence of historic data for statistical estimation (usually for the estimate of variability), hypothesis testing and interval estimation may still be possible if the scientist is prepared to make certain assumptions. For example, suppose that an aquatic biologist is confident that he or she can estimate the variability in IBI in impacted streams based on experience and knowledge of the literature. This estimate could provide the necessary variability measure, but it is obviously conditional on the judgment of the biologist.

None of the approaches presented in this document are without assumptions; even the example in Chapter 2 includes the assumption that the sample data adequately reflect the true situation. Judgment-based estimates of statistics require a different assumption, namely, the assumption that the investigator's judgment is good.

The most serious difficulty in the application of interval estimation and hypothesis testing for biosurvey data is the small sample size associated with many biological surveys. The strength of inferences from statistical analysis is tied to sample size. If expert judgment is not available or not acceptable, then sample size must be large; otherwise, statistical testing is either not possible or not particularly useful. But how large is "large enough"? There is no single, correct answer to that question. As a rule, the stan-

dard error drops according to the square root of the sample size; thus, the answer to the question depends on the error level that is acceptable in the problem under study.

In general, sample sizes greater than 10 are usually desirable, and sample sizes smaller than five may prevent meaningful statistical testing. In addition, since standard error may be expected to drop with the *square root* of the sample size, there are diminishing returns as sample size grows larger.

What can be done when sample size is too small and expert judgment is either not available or not acceptable? Any amount of data or evidence can indicate an effect (or the absence of an effect), and this information can be described in text, presented in tables, or displayed in graphs. However, in the case of very small samples, it is important to emphasize that the analysis is descriptive and not confirmatory. Alternatively, if the investigators have data on biological and chemical indicators of impairment and criteria for each of the indicators, then it may still be possible to test effects across indicators.

Suppose there is no sample size estimate — only an estimate of variability based on expert judgment. How can statistical testing be completed? We actually have some well-established approaches to elicit judgment-based quantities and error estimates, along with an effective number of degrees of freedom (Meyer and Booker, 1991). Alternatively, the scientist may simply summarize test results in a table with sample size (or degrees of freedom) and test results (e.g., *p*-values) given for a range from small to large samples. In some cases, the conclusion may not depend on the effective sample size; in others, sample size may be critical, which places more importance on the goodness of the judgmental assessment.

Assessments Involving Several Indicators

Suppose that sampling has occurred at a stream site at which environmental degradation is suspected, but the sample size for any single indicator is too small for hypothesis testing. For each indicator, the state has established an impairment criterion; thus, the results of sampling could be presented either as a measurement (e.g., dissolved oxygen concentration) or as success or failure in meeting the state's criterion. Each of the indicators is expected to provide an independent measure or assessment of environmental degradation; therefore, several indices cannot be separately included in the analysis if they are based on the same underlying measurements.

As an example, Table 5.1 presents three biological indices, the IBI, ICI, and Iwb based on sampling at a single site on three different dates. The state biocriteria are also given. It is assumed that the two-month period between samplings results in temporal independence between the samples.

Table 5.1—Biological Indices and biocriteria

DATE	IBI	ICI	Iwb
June 15	43(1)	38(1)	9.4(1)
August 15	39(0)	38(1)	8.7(1)
October 15	42(1)	36(1)	8.3(1)
Biocriteria	40	35	8.5

Since we have only a single estimate per date on each index, and only three data points per date and per index, statistical inference opportunities are limited. We can, however, treat the nine index estimates in Table 5.1 as nine independent measures by which to assess the underlying condition of biologic impairment, based on biocriteria violations. The indices in Table 5.1 are recorded as a 0-1 variable, in parentheses, indicating attainment (1) or violation (0) of each biocriterion. Next, these nine 0-1 data points can be subjected to statistical analysis to determine the overall biologic impairment reflected in the aggregate of the three indices.

First, calculate the proportion of violations (p) in the sample as an estimate for the probability of biologic impairment at the site:

$$\hat{p} = \frac{2}{9} = 0.222$$

$\hat{p}$ is a point estimate that is uncertain due to natural variability and measurement error. We can calculate a confidence interval for $\hat{p}$ or test the hypothesis that $\hat{p}$ is less than a specified critical value. Once it is calculated, a confidence interval or a percentile could serve as a cutoff point indicative of biological impairment. For example, one might define impairment as more than 50 percent violations. As a variation on that idea, Rankin and Yoder (1990) selected the 75th percentile in a histogram of sample IBI deviations (from the mean value) to be the limit of tolerable variation.

Confidence intervals for $\hat{p}$ can be determined using binomial tables or graphs like those presented in Hahn and Meeker (1991), or using Table 1.4.1 in Snedecor and Cochran (1967). For example, the two-sided 90 percent confidence interval for this example (based on Table A.23a in Hahn and Meeker) is

$$0.041 \leq p \leq 0.550$$

If instead of binomial tables, the large sample normal approximation is used (see Snedecor and

Cochran, 1967), the two-sided 90 percent confidence interval is

$$\hat{p} - 1.645\sqrt{(\hat{p})(1-\hat{p})/n} \le p \le \hat{p} + 1.645\sqrt{(\hat{p})(1-\hat{p})/n}$$

which, for this example is

$$\frac{2}{9} - 1.645\sqrt{\frac{2}{9}(\frac{7}{9})/9} \le p \le \frac{2}{9} + 1.645\sqrt{\frac{2}{9}(\frac{7}{9})/9}$$

$$0 \le p \le 0.450$$

Clearly, the large sample normal approximation is not appropriate for this small sample.

The binomial confidence interval for $\hat{p}$ is quite large as a consequence of the small sample size; this illustrates how small samples can hamper rigorous statistical inference. Nevertheless, the information in even a small number of samples can be expressed graphically (e.g., using a histogram) or in statistics characterizing center and dispersion. Following Rankin and Yoder, a percentile can be selected from the histogram to serve as a biocriterion.

Note that this percentile reflects variability in the sample, but not strength of evidence as conveyed in sample size or degrees of freedom. The advantage in using a confidence interval rather than an empirical distribution percentile is that the sample size is incorporated in the confidence interval. Thus, more information, expressed as a larger sample size, translates properly to a smaller confidence interval (and indicates greater strength of evidence).

In many applications, intervals may be one-sided, since only one side or bound is of interest. In this example, the two-sided 90 percent confidence interval upper cutoff of 0.55 is the one-sided upper bound on the 95 percent confidence interval. From this information, an infinite number of impairment criteria are possible. One option is to require that $\hat{p} = 0.5$ be outside the upper 95 percent confidence interval for attainment; this could be interpreted as indicating only a slight possibility, a 50/50 chance, of overall biocriteria violation. With that impairment criterion, analysis of the data in Table 5.1 leads to failure to achieve attainment. This conclusion would be reversed, even if the 2/9 biocriteria index violation rate continued, if more samples were collected leading to a tighter confidence interval. The conclusion would also be different for roughly the same sample size if the frequency of biocriteria index violation were lower.

Regional Reference Data

Bioassessment data on regional conditions (e.g., regional reference sites) may sometimes be used with small sample sizes, or even with a single sample, to go beyond a point estimate of status. Consider, for example, the information presented in Yoder (1991). He compared the assessments from the application of the Ohio narrative macroinvertebrate criteria from 1979 through 1986 with a calculated ICI score. In this study, about 400 sites were rated using both narrative macroinvertebrate criteria and calculated ICI; and the two ratings were then compared for each of the sites.

Yoder expressed this comparison using three ICI distributions: the ICI scores for the sites labeled "good/exceptional" based on the narrative criteria, the ICI scores for the "fair" sites, and the ICI scores for the "poor/very poor" sites (see Yoder, 1991, Fig. 7). Yoder argued that the ICI scores are more reliable than are the classifications based on the narrative criteria, and he employed point cutoffs between classes (ICI = 35 between good and fair; ICI = 13 between fair and poor).

If, unlike Yoder, we take the ICI distributions for each of the three classes as reference distributions, then we can use the classification rules typically employed with discriminant analysis (see Flury and Riedwyl, 1988) to estimate the probability that any new sites belong in each class. To do this, we must make a distributional assumption concerning the probability model that describes the ICI within each class. As a rule, it is assumed that this distribution (of ICI) is normal (within each of the three classes), with mean and variance estimated based on the sample (ICI) values.

As another example of small sample size data sets, imagine that repeated IBI measurements are taken both from a reference site, and from a site with known anthropogenic pollutants. Data from each of the sites are analyzed, and their respective distribution functions are created. Such a case is presented in Figure 5.1. Here, the IBI sampling distributions for each site are roughly shaped as a normal distribution.

Assume, further, that a single IBI measurement from a third site is generated. This single measurement is shown on Figure 5.1 as a solid vertical line. Does the investigator have enough information to categorize the site as impacted or not impacted? Visual examination of the figure shows that the third site IBI measurement lies in an area of substantial overlap between the impacted and reference site distributions. Therefore, given that the sampling error of the third site is unknown, it is difficult to assess with confidence whether the third IBI measurement is consistent with either the reference or impacted sites.

At this point, the investigator would be best served if he or she gathered additional IBI measurements. If, on the other hand, the single-sample IBI measurement had been in the tail of either distribution (say, an IBI of 20 or 55), then the investigator could have classified the third site appropriately. In

making this classification, the investigator would have noticed that little overlap of the distributions occurs in the extreme tails of the impacted and reference site distributions.

Using Background Variability Measures

In the previous section, the Ohio ICI biocriteria were identified as point values between classes (e.g., ICI = 35 is the warmwater habitat criterion separating "good/exceptional" from "fair"). When a single ICI determination is available from a new site, the Ohio criteria can be used to classify the site, ignoring uncertainty. Beyond that, if it is assumed that the Ohio ICI classification scheme is fixed and certain, and if a reliable estimate of site ICI variability is available, then the classification based on a single ICI value can be assessed using a hypothesis test.

In situations with only a single estimate of a bioindicator, collateral information must be obtained to provide the estimate of variability. There are several potential measures of site bioindicator variability that might be suitable; Rankin and Yoder's (1990) discussion presents several informative graphs to show, for example, that the IBI coefficient of variation drops as IBI increases (Rankin and Yoder, 1990, Fig. 2), and IBI coefficient of variation increases slightly as drainage area increases (ibid., Fig. 7).

Knowledge and judgment can be quite helpful in selecting the variability estimate. For example, if it is believed that the site bioindicator variability is roughly constant within a specified category, then a calculated estimate of variability for the bioindicator within the appropriate class can be used as the variability measure for the site of interest. Categories may be selected on any criterion (e.g., ecoregion, IBI range) that is scientifically plausible and leads to an acceptably large overall sample size for variability estimation.

Rankin and Yoder's graphs suggest that, while the IBI coefficient of variation changes with selected categories (IBI range), the IBI standard deviation may be roughly constant across IBI classes and across ecoregions. A median standard deviation between 4 and 5 appears to be quite consistent in the graphs. Based on this collateral information, it is assumed that site-specific IBI in Ohio, under constant conditions (i.e., no change in site factors that determine IBI), has a standard deviation of 4.5.

Here is an example of how this estimate is used. Assume that the single IBI measurement shown in Figure 5.1 (IBI = 35) was taken in Ohio under the conditions described. Since the sampling program in Ohio is quite large, 4.5 is effectively the true standard deviation for all sites; thus, with a single sample, it may be concluded that the standard error for the mean value (IBI = 35) is also 4.5. To determine whether the sample is taken from the reference or impacted distribution, assume that 18 IBI samples were taken at the reference and impacted sites, and that the following statistics are calculated:

Reference site sample mean = 42, sample standard deviation = 5;

Impacted site sample mean = 27, sample standard deviation = 8.

Then, a two-tailed t test using Equation 2.1b (see Chapter 2) evaluating the null hypothesis that the means are the same will result in the following:

t = 1.43, for the hypothesis that the reference site mean is equal to the mean of the third site mean; and

t = 1.245, for the hypothesis that the impacted site mean is equal to the third site mean.

Based on this information, the investigator has some evidence that the sample collected from the third site is closer to the impacted site mean than to the reference site mean. However, as conveyed by the similar t statistic results, the confidence in this conclusion is relatively weak.

Final Suggestions for Small Sample Sizes

The discussion and examples in this chapter, while intended as useful, general guidance, are not firmly rooted in statistical theory and hence not always to be followed. Rather, they reflect our experience and observations. Further, they concern the real situations that biologists confront — situations that do not conform to well-established statistical procedures. However difficult and awkward for statistical analysis, the problems must be addressed. With this caveat, the following concluding comments summarize the discussion and examples presented here:

1. If the sample size is 1, a measure of variability may still be obtained using expert judgment or other data. If no variability measure can be justified, then descriptive statistics may be the extent of the analysis (i.e., no interval estimation or hypothesis testing).

2. If the sample size is more than 1 but still small (perhaps 5 or fewer), then it is possible to use the sample to estimate variability for interval estimation or hypothesis testing. However, the intervals may be very large and the tests not

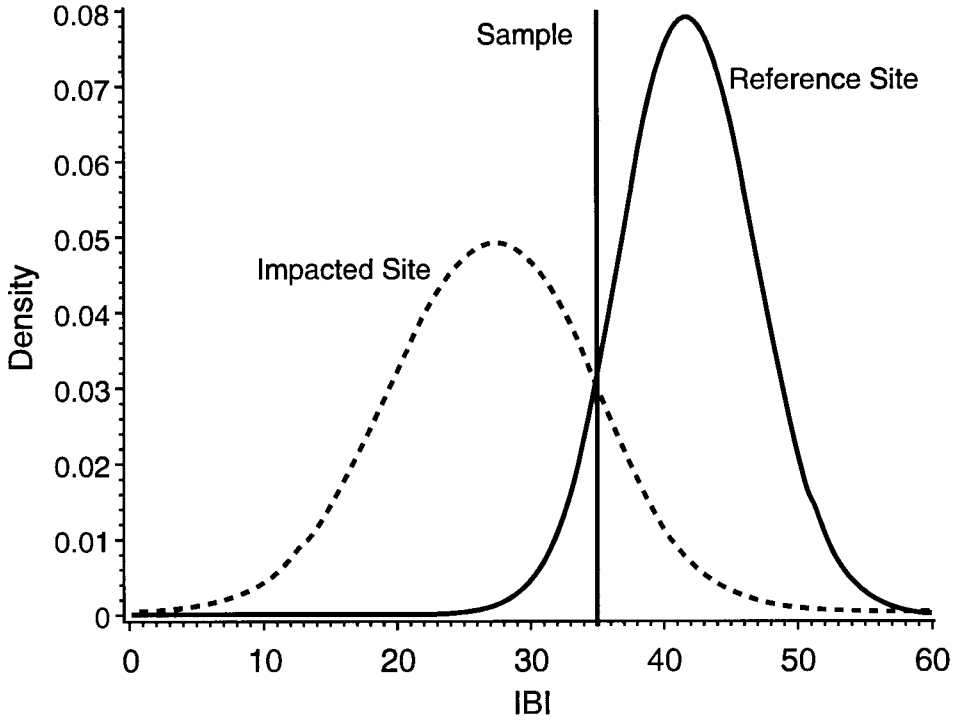

Figure 5.1—IBI Distributions for reference and impacted sites

very powerful, because small sample size means that the strength of evidence is weak.

3. Situations may exist with more than a single estimate of variability. Perhaps one estimate will be based on data and a second estimate on expert judgment. In that case, the two estimates of variance can be pooled, using an estimator like that in Chapter 4's "Reference Distribution Based on Random Sampling Model, Internal Value for σ." A difficulty in pooling when a judgmental estimate of variance is involved is determination of the degrees of freedom for the judgmental variance estimate. Perhaps the best approach is to make a reasoned guess as to how much information the judgment contains with respect to samples (the "effective sample size"):

   (a) if the judgment is highly uncertain, assign it a small number of degrees of freedom (perhaps 2-5),

   (b) if there is more confidence in the judgment, assign the judgment estimate 5+ degrees of freedom.

   If the conclusions from this analysis are not particularly sensitive to the exact choice of the effective sample size for the judgmental estimate, then inferences may be made with some confidence. If, however, the conclusions are sensitive to this choice, then the best approach may be to obtain additional information before drawing final conclusions.

Decision Analysis and Uncertainty

In the preliminary approach presented here we have advocated the use of classical statistical hypothesis testing to summarize data concerning biological criteria. We assume that a decision and succinct conclusions based on the data are needed. However, alternatives to hypothesis testing may be appropriate in certain situations. For example, statistical and graphic summaries (e.g., confidence intervals, bivariate plots) may be used to summarize and present information when the investigator believes that a classical hypothesis test based on a single parameter is too brief or that more evidence should be presented.

An alternative is to recast the hypothesis testing problem using a decision analytic framework. Decision analysis (Raiffa, 1968; Reckhow, 1984) begins with the scientific base summarized in the hypothesis test and incorporates the consequences (e.g., costs and benefits) of possible decisions. In an informal analysis, a decision analytic approach may be undertaken by the decision maker if a desired outcome of management action is "to hedge away" from large adverse consequences or losses. Informal considerations and hedging may be most effectively undertaken in an a priori assessment of costs and

benefits, which then becomes a primary basis for choosing between various levels of test significance. Thus, if it seems likely that biological degradation can be avoided, then the decision maker may request that the biologist set the significance level for testing (e.g., that H_0 has no impact) relatively high (e.g., at 0.10 or 0.20). Alternatively, if cleanup costs are high relative to benefits, then the test significance level (for H_0 has no impact) could be set relatively low (e.g., at 0.01 or 0.005).

Suppose that a measure of biological integrity is tested for upstream-downstream differences surrounding wastewater treatment plant discharges from small treatment plants (less than 5 million gallons per day) throughout the state. If the per person cost to upgrade the treatment level for small communities is generally quite high, and the benefits to be derived from biological improvements are generally low (relative to the organisms affected and typical uses of the streams), hedging away from high cost may be informally undertaken by setting the significance (or "action") level of the test quite low (e.g., 0.01 or 0.005). Additional study of biological degradation, costs, and benefits would be triggered only if an upstream-downstream test result was significant at this level.

Hedging away from large losses is an option precisely because of scientific uncertainty. If there were no scientific uncertainty about biological degradation, then the analysis would always focus on costs and benefits, and the management option with the highest net benefits would be selected. On the other hand, if scientific uncertainty is extreme, an appropriate strategy may be either to hedge farther from large adverse consequences or to seek more information, if possible, to reduce scientific uncertainty before new management action is adopted.

In more formal applications, decision analysis may be used to combine uncertain scientific information on biocriteria (expressed probabilistically) with an overall measure of net benefits or use associated with management actions. This approach is most effective in a Bayesian context; Reckhow (1984) presents a simple example applied to lake eutrophication management. However, comprehensive Bayesian decision analysis is apt to be prohibitively expensive (in terms of human resources and cost) for all but the most critical and consequential problems.

One outcome of data analysis may be that the decision maker will desire more information before implementing new management actions. In formal decision analysis, a value of information calculation should be made to help one determine the wisdom of immediate action versus additional data collection and analysis. In informal analysis, one should consider how useful new information would be if action has to be deferred pending its arrival.

The outcome of hypothesis testing is a statistical summary of evidence on biological degradation. It does not establish cause and effect, although a well-designed test may associate degradation with a candidate cause. The strength of causal conclusions depends on a number of factors including a priori scientific knowledge and field observation. Scientific support for management actions is greatest when the observation of degradation is accompanied by documentation of a causal relationship.

In most cases, environmental management decisions reflect a certain limited understanding of causal connections and a certain degree of observational evidence that is more statistical in nature. This combination is a reasonable basis for decision; in fact, it would be unreasonable to expect detailed causal knowledge in support of every decision. However, as management actions are undertaken and biological response is observed after the fact, more observational evidence may be gathered to support earlier decisions.